บทที่ 5

การตัดคำโดยใช้พจนานุกรมภาษาไทย

บทนี้จะกล่าวถึงการนำพจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่ออกแบบโครงสร้างข้อมูลมาใช้ประโยชน์ในการตัดคำ ซึ่งมีรายละเอียดของอัลกอริทึมการตัดคำ และตัวอย่างการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่อธิบายรูปแบบโครงสร้างไว้แล้วในบทที่ 4

อัลกอริทึมการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

อัลกอริทึมการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทยที่จะเสนอในวิทยานิพนธ์ครั้งนี้ มีจุดเริ่มต้นจากการค้นหาคำศัพท์จากประโยคที่ต้องการตัดคำ แล้วนำคำศัพท์เหล่านั้นจัดเก็บไว้ในอะเรย์ชุดหนึ่งจากนั้นจึงสร้างอะเรย์ของคำศัพท์ชุดต่อไป โดยค้นหาคำศัพท์จากประโยคอินพุทที่ได้จากการตัดคำที่ค้นพบในพจนานุกรม แต่ทั้งนี้พจนานุกรมฯ จะต้องบรรจุในหน่วยความจำหลักตลอดเวลา

ก่อนที่จะอธิบายอัลกอริทึมการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทยนั้น จะแสดงตัวอย่างการตัดคำโดยใช้พจนานุกรมฯ รวมทั้งลักษณะโครงสร้างประโยค เพื่อให้สามารถเข้าใจอัลกอริทึมการตัดคำได้รวดเร็วยิ่งขึ้น

1. ตัวอย่างการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย

สมมติว่าเราจัดเก็บคำศัพท์ต่อไปนี้ กา การ มอบ รม รางวัล อบ ไว้ในพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย แล้วต้องการตัดคำในประโยค "การมอบรางวัล" โดยใช้พจนานุกรม จะมีขั้นตอนแสดงไว้ในตารางที่ 5.1

2. ลักษณะโครงสร้างประโยค

ลักษณะโครงสร้างประโยคที่จะอธิบายในบทนี้จะใช้สัญญลักษณ์ต่างๆ ช่วยในการอธิบายดังนี้

ST แทนประโยค

S_t แทนเซ็ทของคำ

i แทนลำดับของเซ็ท

$S_1, S_2, \ldots, S_n$ เป็นเซ็ทของคำที่ปรากฏในประโยค เป็นลำดับที่ $1, 2, \ldots, n$

S_t แทนจำนวนสมาชิกของเซ็ท S_t

w_{ij} แทนคำศัพท์ (สมาชิกในเซ็ท S_t ลำดับที่ j)

j แทนลำดับที่ของคำศัพท์ในเซ็ท

รูปแบบของประโยคโดยทั่วไปประกอบด้วยชุดของคำศัพท์ที่เรียงติดต่อกันอย่างมีกฎเกณฑ์ ซึ่งอาจเขียนแสดงด้วยสัญญลักษณ์ทางคณิตศาสตร์ได้ดังนี้

$$ST = S_1, S_2, S_3 \ldots, S_{n-1}, S_n$$

นั่นคือหากจะสร้างประโยคจากเซ็ทของคำ $S_1, S_2, S_3 \ldots, S_{n-1}, S_n$ แล้วจำนวนประโยคที่สามารถสร้างได้จะน้อยกว่าหรือเท่ากับ $S_1 * S_2 * S_3 \ldots S_{n-1} * S_n$ ประโยคหรือแทนได้ด้วยสัญญลักษณ์

$$\prod_{i=1}^{n} S_i$$

สำหรับตัวอย่างของการสร้างประโยคจากเซ็ทของคำมีดังนี้เช่น ประโยค "นอนตากลม" เมื่อพิจารณาแล้วพบว่าสามารถแยกคำศัพท์เป็น 2 คำ

คำที่ 1 นอน

คำที่ 2 ตากลม ซึ่งเมื่อพิจารณาแล้วพบว่าสามารถแยกเป็น ตา กลม หรือ ตาก ลม

และเมื่อนำคำศัพท์มาสร้างเป็นประโยคจะได้ 2 ประโยค คือ

1 นอน ตา กลม

2 นอน ตาก ลม

อีกตัวอย่างก็คือ "เรือโคลงเพราะโคลงเรือ" หากเราพิจารณาประโยคดังกล่าวแล้วพบว่าคำศัพท์แยกออกเป็น 5 ชุด ดังนี้

คำที่ 1 เรือ

คำที่ 2 โคลง ซึ่งเมื่อพิจารณาอีกครั้งแล้วจะแยกได้เป็น 2 คำ คือ โคลง หรือ โค ลง

คำที่ 3 เพราะ

คำที่ 4 โคลง ซึ่งเมื่อพิจารณาอีกครั้งแล้วจะแยกได้เป็น 2 คำ คือ โคลง หรือ โค ลง

คำที่ 5 เรือ

เมื่อนำคำศัพท์ทั้งหมดมาสร้างเป็นประโยคจะได้ 4 ประโยค คือ

1 เรือ โคลง เพราะ โคลง เรือ

2 เรือ โคลง เพราะ โค ลง เรือ

3 เรือ โค ลง เพราะ โคลง เรือ

4 เรือ โค ลง เพราะ โค ลง เรือ

แต่ประโยคที่ให้ความหมายถูกต้องมี 2 ประโยค คือประโยคที่ 1 และ 2

จากตัวอย่างการสร้างประโยคทั้ง 4 ตัวอย่างพบว่าจำนวนประโยคที่สร้างขึ้นในตัวอย่างแรกที่เป็นไปได้คือ 1*2 = 2 และจำนวนประโยคที่สร้างได้จริงคือ 2 ประโยคซึ่งเท่ากับจำนวนประโยคที่เป็นไปได้

ส่วนตัวอย่างที่ 2 จำนวนประโยคที่เป็นไปได้ที่สร้างขึ้นเท่ากับ 1*2*1*2*1 = 4 และจำนวนประโยคที่สร้างได้และให้ความหมายถูกต้องมี 2 ประโยค ซึ่งน้อยกว่าจำนวนประโยคที่สร้างได้จริง

จะเห็นได้ว่าจำนวนประโยคที่สร้างขึ้นจากเซ็ทของคำน้อยกว่าหรือเท่ากับผลคูณของจำนวนคำศัพท์ที่เกิดขึ้นจริง

สำหรับลักษณะโครงสร้างของประโยคอาจแสดงตามรูปที่ 5.1

3. <u>อัลกอริทึมการตัดคำโดยใช้พจนานุกรมอิเล็กทรอนิกส์ภาษาไทย</u>

อัลกอริทึมการตัดคำที่นำเสนอไว้ในวิทยานิพนธ์ฉบับนี้ เป็นอัลกอริทึมที่ปรับมาจากอัลกอริทึมการค้นหาคำศัพท์และลักษณะโครงสร้างของประโยค ซึ่งการสร้างอะเรย์

ของคำศัพท์เป็นการค้นหาคำศัพท์ที่เป็นไปได้ทั้งหมดที่ขึ้นต้นด้วยส่วนของพยัญชนะที่เหมือนกันที่จัดเก็บไว้ในพจนานุกรมฯ โดยปฏิบัติกับประโยคเสมือนคำศัพท์ที่ต้องการค้น และผลลัพธ์ที่ได้เป็นอะเรย์ที่จัดเก็บความยาวของคำศัพท์ที่ปรากฏในพจนานุกรม จากนั้นจึงนำความยาวมาสร้างเป็นคำศัพท์เพื่อพิจารณาในขั้นต่อไปว่า เมื่อดึงคำศัพท์นั้นออกจากประโยคแล้วสามารถสร้างอะเรย์ของคำศัพท์จากประโยคส่วนที่เหลือได้หรือไม่ หากคำศัพท์คำใดไม่สามารถสร้างอะเรย์ได้ต่อไป แสดงว่าการแยกคำศัพท์ที่ตำแหน่งนั้นยังไม่ถูกต้อง ให้สร้างอะเรย์ใหม่โดยใช้คำศัพท์คำต่อไปของอะเรย์ หรืออาจสรุปขั้นตอนง่ายๆ ก็คือ สร้างอะเรย์เก็บชุดคำศัพท์ทั้งหมดจากประโยคอินพุท แล้วนำคำศัพท์ในอะเรย์เหล่านั้นมาเรียบเรียงเป็นประโยคใหม่โดยที่ผลลัพธ์ที่ได้จะเป็นประโยคที่ได้ตัดคำตามคำศัพท์ที่ค้นหาได้จริงจากพจนานุกรม

ขั้นตอนการตัดคำที่กล่าวถึงในวิทยานิพนธ์นี้อาจสรุปเป็นข้อๆ ได้ดังนี้

ก. สร้างอะเรย์ของคำศัพท์ที่ค้นหาจากพจนานุกรมฯ ของประโยคอินพุท

ข. ดึงคำศัพท์ในอะเรย์ชุดแรกมาแล้วสร้างอะเรย์ของคำศัพท์ที่ค้นหาได้จากพจนานุกรมฯ ของประโยคอินพุทประโยคใหม่ (นำคำศัพท์ที่ดึงมาจากอะเรย์ชุดแรกตัดออกจากประโยคอินพุทประโยคแรก)
กระทำดังนี้เรื่อยไปจนกระทั่งสิ้นสุดประโยคอินพุท และให้อะเรย์ของคำชุดสุดท้ายที่สร้างขึ้นเป็นอะเรย์ที่ n

ค. ย้อนการทำงานกลับมาที่อะเรย์ชุดที่ n-1 โดยนำคำศัพท์ในอะเรย์ชุดที่ n-1 ที่ยังไม่ได้ ใช้งานมาสร้างอะเรย์ของคำศัพท์ชุดที่ n ใหม่ และใช้คำศัพท์ในอะเรย์ชุดที่ n-1 มาสร้างอะเรย์ของคำศัพท์ชุดใหม่เรื่อยๆ จนกระทั่ง คำศัพท์ในอะเรย์ชุดที่ n - 1 หมด

ง. เมื่อคำศัพท์ในอะเรย์ชุดที่ n-1 หมดให้ย้อนการทำงานกลับมาอะเรย์ชุดที่ n-2 แล้ว เลือกคำศัพท์ในอะเรย์ชุดที่ n-2 คำที่ยังไม่ถูกเลือกต่อจากนั้นสร้างอะเรย์คำศัพท์อะเรย์ที่ n-1 ใหม่ แล้วเลือกคำศัพท์ในอะเรย์ที่ n-1 เพื่อนำไปสร้างอะเรย์คำศัพท์ในอะเรย์ n ใหม่

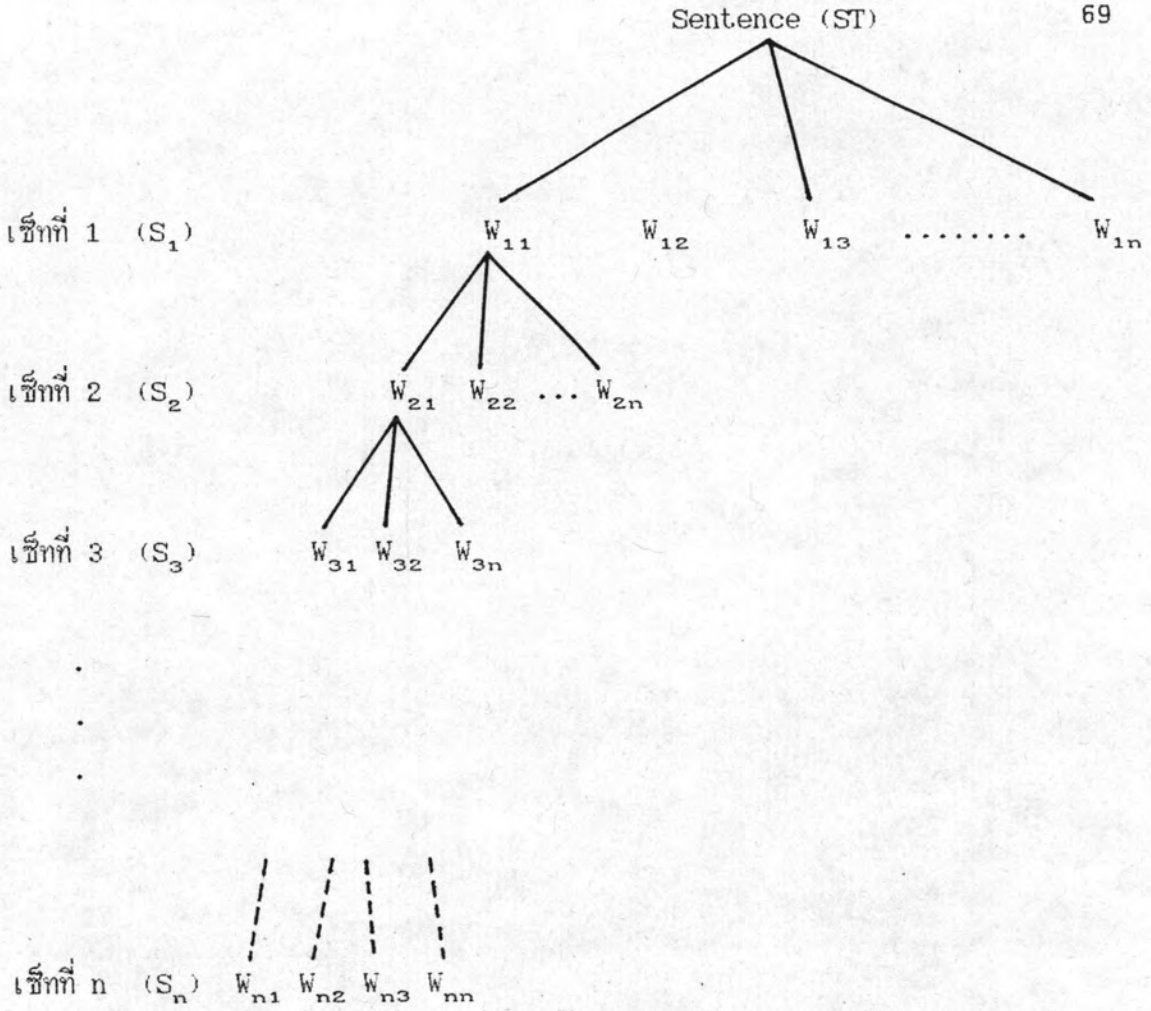
จ. ย้อนการทำงานกลับมาที่อะเรย์ชุดก่อนหน้าเรื่อยไปจนกระทั่งถึงอะเรย์ชุดที่ 1 และคำศัพท์ในอะเรย์ชุดที่ 1 หมด

	อินพุท	ขั้นตอนการปฏิบัติการ	เอาท์พุท
1	"การมอบรางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ กา ,การ
2	"กา"	นำคำศัพท์จากอะเรย์ที่ 1 มาใช้	"รมอบรางวัล"
	"การมอบรางวัล"	แล้วตัดคำออกจากประโยค	
3	"รมอบรางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ รม
4	"รม"	ดึงคำศัพท์จากอะเรย์ในข้อ 2	"อบรางวัล"
	"รมอบรางวัล"	มาใช้ แล้วตัดคำนั้นออก	
5	"อบรางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ อบ
6	"อบ"	ดึงคำศัพท์จากอะเรย์ในข้อ 2	"รางวัล"
	"อบรางวัล"	มาใช้ แล้วตัดคำนั้นออก	
7	"รางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ รางวัล
8		ตรวจสอบว่าสิ้นสุดประโยคแล้ว	กา รม อบ รางวัล
		แสดงประโยคที่มีการตัดคำ	
9	"การ"	ถอยกลับไปดึงคำศัพท์จากอะเรย์	"มอบรางวัล"
	"การมอบรางวัล"	ในข้อ 7 มาปรากฏว่าคำศัพท์ใน	
		อะเรย์ดังกล่าวหมดแล้วจึงถอย	
		กลับไปดึงคำศัพท์ในอะเรย์ก่อน	
		หน้า(ข้อ5,3) ปรากฏว่า	
		คำศัพท์จากอะเรย์ของทั้ง 2 ข้อ	
		หมดแล้วจึงถอยไปดึงคำศัพท์จาก	
		อะเรย์ในข้อ 1 มาใช้	
10	"มอบรางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ มอบ
11	"มอบ"	นำคำศัพท์คำแรกในเซ็ทมาใช้	"รางวัล"
	"มอบรางวัล"	แล้วตัดคำนั้นออกจากประโยค	
12	"รางวัล"	ค้นหาคำศัพท์ในพจนานุกรมฯ	อะเรย์คำศัพท์ รางวัล

ตารางที่ 5.1 แสดงขั้นตอนการตัดคำโดยใช้พจนานุกรม

	อินพุท	ขั้นตอนการปฏิบัติการ	เอาท์พุท
13		พบว่าสิ้นสุดคำในประโยค แสดงประโยคที่มีการตัดคำ	การ มอบ รางวัล
14		ถอยไปดึงคำศัพท์ในอะเรย์ที่ 1 ปรากฏว่าคำศัพท์หมด แสดงผลลัพธ์ออกมา	

ตารางที่ 5.1 (ต่อ) แสดงขั้นตอนการตัดคำโดยใช้พจนานุกรม



รูปที่ 5.1 แสดงลักษณะโครงสร้างของประโยค

<u>การพัฒนาโปรแกรมการตัดคำ</u>

สำหรับโปรแกรมการตัดคำนั้น ได้พัฒนาโปรแกรมการตัดคำสำหรับพจนานุกรมที่มีโครงสร้างแบบดับเบิลอะเรย์ และโปรแกรมการตัดคำสำหรับพจนานุกรมที่มีโครงสร้างตามที่ให้นิยามไว้ในบทที่ 4 ด้วย